Dynamic models 2
Switching KFs continued,
Assumed density filters,
DBNs, BK, extensions

Probabilistic Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University

November 21st, 2005
Announcement

- Special recitation lectures
  - Pradeep will give two special lectures
  - Nov. 22 & Dec. 1: 5-6pm, during recitation
  - Covering: variational methods, loopy BP and their relationship
  - Don’t miss them!!!

- It’s FCE time!!!
  - Fill the forms online by Dec. 11
  - [www.cmu.edu/fce](http://www.cmu.edu/fce)
  - It will only take a few minutes
  - Please, please, please help us improve the course by providing feedback
Last week in “Your BN Hero”

- Gaussian distributions reviewed
  - Linearity of Gaussians
  - Conditional Linear Gaussian (CLG)

- Kalman filter
  - HMMs with CLG distributions
  - Linearization of non-linear transitions and observations using numerical integration

- Switching Kalman filter
  - Discrete variable selects transition model depends
  - Mixture of Gaussians represents belief state
  - Number of mixture components grows exponentially in time
The moonwalk
Last week in “Your BN Hero”

- Gaussian distributions reviewed
  - Linearity of Gaussians
  - Conditional Linear Gaussian (CLG)

- Kalman filter
  - HMMs with CLG distributions
  - Linearization of non-linear transitions and observations using numerical integration

- Switching Kalman filter
  - Discrete variable selects transition model depends
  - Mixture of Gaussians represents belief state
  - Number of mixture components grows exponentially in time
Switching Kalman filter

- At each time step, choose one of $k$ motion models:
  - You never know which one!
- $p(X_{i+1}|X_i, Z_{i+1})$
  - CLG indexed by $Z_i$
  - $p(X_{i+1}|X_i, Z_{i+1}=j) \sim N(\beta^j_0 + B^j X_i; \Sigma^j_{X_{i+1}|X_i})$
Inference in switching KF – one step

- Suppose
  - p(X₀) is Gaussian
  - Z₁ takes one of two values
  - p(X₁|X₀,Z₁) is CLG

- Marginalize X₀
  \[ p(x₁|z₁) = \int_{x₀} p(x₀) \cdot p(x₁|x₀, z₁) \, dx₀ \]

- Marginalize Z₁
  \[ p(x₁) = \sum_j p(x₁|z₁=j) \cdot p(z₁=j) \]

- Obtain mixture of two Gaussians!
Multi-step inference

- Suppose
  - $p(X_i)$ is a mixture of $m$ Gaussians
  - $Z_{i+1}$ takes one of two values
  - $p(X_{i+1}|X_i, Z_{i+1})$ is CLG

- Marginalize $X_i$

- Marginalize $Z_i$

- Obtain mixture of $2m$ Gaussians!

- Number of Gaussians grows exponentially!!

$$p(x_i) = \sum_{k=1}^{m} w_k \mathcal{N}(\mu_k, \Sigma_k)$$

$$p(x_{i+1}|z_{i+1} = j) = \int_{x_i} p(x_{i+1}|x_i, z_{i+1} = j) \cdot p(x_i) \, dx_i$$

$$= \sum_{k=1}^{m} w_k \int_{x_i} p(x_{i+1}|x_i, z_{i+1} = j) \mathcal{N}(\mu_k, \Sigma_k) \, dx_i$$

$$p(x_{i+1}) = \sum_{j} p(z_{i+1} = j) \cdot p(x_{i+1}|z_{i+1} = j)$$
Visualizing growth in number of Gaussians
Computational complexity of inference in switching Kalman filters

- Switching Kalman Filter with (only) 2 motion models

- Query: $p(x_n)$

- Problem is NP-hard!!! [Lerner & Parr `01]
  - Why “!!!”? Graphical model is a tree:
    - Inference efficient if all are discrete
    - Inference efficient if all are Gaussian
    - But not with hybrid model (combination of discrete and continuous)
Bounding number of Gaussians

- $P(X_i)$ has $2^m$ Gaussians, but...
- usually, most are bumps have low probability and overlap:

Intuitive approximate inference:
- Generate $k.m$ Gaussians
- Approximate with $m'$ Gaussians
Collapsing Gaussians – Single Gaussian from a mixture

- Given mixture $P \sim \mathcal{N}(\mu, \Sigma)$
- Obtain approximation $Q \approx \mathcal{N}(\mu, \Sigma)$ as:
  
  $$
  \mu = \sum_i w_i \mu_i \quad \text{weighted sum of } \mu_i \\
  \Sigma = \sum_i w_i \Sigma_i + \sum_i w_i (\mu_i - \mu)(\mu_i - \mu)^T \\
  \text{approximate weighted sum of } \Sigma_i
  $$

- Theorem:
  - $P$ and $Q$ have same first and second moments
  - KL projection: $Q$ is single Gaussian with lowest KL divergence from $P$

$$
Q = \arg\min_{Q \sim \mathcal{N}} \text{KL}(P \parallel Q)
$$
Collapsing mixture of Gaussians into smaller mixture of Gaussians

- Hard problem!
  - Akin to clustering problem…
    - Similar to fitting mixture of $K$-Gaussians to data

- Several heuristics exist
  - c.f., Uri Lerner’s Ph.D. thesis
Operations in non-linear switching
Kalman filter

- Compute mixture of Gaussians for $p(X_t \mid O_1:t = o_1:t)$
- Start with $p(X_0)$
- At each time step $t$:
  - For each of the $m$ Gaussians in $p(X_{0:t})$:
    - **Condition** on observation (use numerical integration)
    - **Prediction** (Multiply transition model, use numerical integration)
      - Obtain $k$ Gaussians
    - **Roll-up** (marginalize previous time step)
  - **Project** $k.m$ Gaussians into $m'$ Gaussians $p(X_{0:t+1})$
Assumed density filtering

Examples of very important assumed density filtering:

- Non-linear KF
- Approximate inference in switching KF

General picture:

- Select an assumed density
  - e.g., single Gaussian, mixture of $m$ Gaussians, ...
- After conditioning, prediction, or roll-up, distribution no-longer representable with assumed density
  - e.g., non-linear, mixture of $k.m$ Gaussians,…
- Project back into assumed density
  - e.g., numerical integration, collapsing,…
When non-linear KF is not good enough

- Sometimes, distribution in non-linear KF is not approximated well as a single Gaussian
  - e.g., a banana-like distribution

- Assumed density filtering:
  - Solution 1: **reparameterize problem** and solve as a **single Gaussian**
  - Solution 2: more typically, **approximate as a mixture of Gaussians**
Distributed Simultaneous Localization and Tracking

- Place cameras around an environment, don’t know where they are
- Could measure all locations, but requires lots of grad. student time
- Intuition:
  - A person walks around
  - If camera 1 sees person, then camera 2 sees person, learn about relative positions of cameras
Donut and Banana distributions

- Observe person at distance $d$
- Camera could be anywhere in a ring
Gaussians represent “balls”

- Gaussian approximation leads to poor results
- Can’t apply standard Kalman filter 😞
- Or can we… 😊
Reparameterized KF for SLAT
Example of KF – SLAT
Simultaneous Localization and Tracking
When a single Gaussian ain’t good enough

- Sometimes, smart parameterization is not enough
  - Distribution has multiple hypothesis

- Possible solutions
  - Sampling – particle filtering
  - Mixture of Gaussians
  - …

- Quick overview of one such solution…

[Fox et al.]
Approximating non-linear KF with mixture of Gaussians

- Robot example:
  - P(X_i) is a Gaussian, P(X_{i+1}) is a banana
  - Approximate P(X_{i+1}) as a mixture of m Gaussians
    - e.g., using discretization, sampling,…
  - Problem:
    - P(X_{i+1}) as a mixture of m Gaussians
    - P(X_{i+2}) is m bananas
  - One solution:
    - Apply collapsing algorithm to project m bananas in m’ Gaussians
What you need to know about switching Kalman filters

- **Kalman filter**
  - Probably most used BN
  - Assumes Gaussian distributions
  - Equivalent to linear system
  - Simple matrix operations for computations

- **Non-linear Kalman filter**
  - Usually, observation or motion model not CLG
  - Use numerical integration to find Gaussian approximation

- **Switching Kalman filter**
  - Hybrid model – discrete and continuous vars.
  - Represent belief as mixture of Gaussians
  - Number of mixture components grows exponentially in time
  - Approximate each time step with fewer components

- **Assumed density filtering**
  - Fundamental abstraction of most algorithms for dynamical systems
  - Assume representation for density
  - Every time density not representable, project into representation
More than just a switching KF

- Switching KF selects among \( k \) motion models
- Discrete variable can depend on past
  - Markov model over hidden variable

What if \( k \) is really large?
  - Generalize HMMs to large number of variables
Dynamic Bayesian network (DBN)

HMM defined by
- Transition model $P(X_{t+1}|X_t)$
- Observation model $P(O_t|X_t)$
- Starting state distribution $P(X_0)$

DBN – Use Bayes net to represent each of these compactly
- Starting state distribution $P(X_0)$ is a BN
- (silly) e.g., performance in grad. school DBN
  - Vars: Happiness, Productivity, Hirability, Fame
  - Observations: Paper, Schmooze

$P(X_0)$: $H_0 \quad P_0 \quad B_0 \quad F_0$

$P(X_{t+1}|X_t)$ $P(O_{t+1}|X_{t+1})$
Transition Model: Two Time-slice Bayes Net (2-TBN)

- Process over vars. $X$

- 2-TBN: represents transition and observation models $P(X_{t+1}, O_{t+1} | X_t)$
  - $X_t$ are *interface variables* (don’t represent distribution over these variables)
  - As with BN, exponential reduction in representation complexity

![Diagram](attachment:diagram.png)
Unrolled DBN

- Start with $P(X_0)$
- For each time step, add vars as defined by 2-TBN
“Sparse” DBN and fast inference
Almost!
BK Algorithm for approximate DBN inference  
[Boyen, Koller ’98]

- Assumed density filtering:
  - Choose a factored representation $\hat{b}$ for the belief state
  - Every time step, belief not representable with $\hat{b}$, project into representation

Time $\rightarrow t \rightarrow t+1 \rightarrow t+2 \rightarrow t+3$

Diagrams showing the belief state transitions and projected beliefs.
Computing factored belief state in the next time step

- Introduce observations in current time step
  - Use J-tree to calibrate time $t$ beliefs

- Compute $t+1$ belief, project into approximate belief state
  - marginalize into desired factors
  - corresponds to KL projection

- Equivalent to computing marginals over factors directly
  - For each factor in $t+1$ step belief
    - Use variable elimination
Error accumulation

- Each time step, projection introduces error
- Will error add up?
  - causing unbounded approximation error as $t \to \infty$
Contraction in Markov process

At time $t$

$p$:diff. high

$\Delta t$

diffusion after transition

mixing rate $\gamma < 1$

$t+1$

$d(p_t,q_t) > d(p_{t+1},q_{t+1})$

overlap
BK Theorem

- Error does not grow unboundedly!

\[
d(p_2; q_2) = d(Tp_1, Tq_1) + d(Tq_1, q_2) \leq \gamma d(p_1, q_1)
\]
Example – BAT network  [Forbes et al.]
BK results [Boyen, Koller ’98]

Typical evolution of error

Comparing partitions
Thin Junction Tree Filters [Paskin ‘03]

- BK assumes fixed approximation clusters
- TJTF adapts clusters over time
  - attempt to minimize projection error
Hybrid DBN (many continuous and discrete variables)

- DBN with large number of discrete and continuous variables
- # of mixture of Gaussian components blows up in one time step!
- Need many smart tricks…
  - e.g., see Lerner Thesis

Reverse Water Gas Shift System (RWGS) [Lerner et al. ’02]
DBN summary

- **DBNs**
  - factored representation of HMMs/Kalman filters
  - sparse representation does not lead to efficient inference

- **Assumed density filtering**
  - BK – factored belief state representation is assumed density
  - Contraction guarantees that error does blow up (but could still be large)
  - Thin junction tree filter adapts assumed density over time
  - Extensions for hybrid DBNs